Improving Pointwise Predictions with a Pairwise Ranker

Eric Hallman[1]

[1]eBay, 2025 Hamilton Ave, San Jose, CA, 95125, USA

Abstract

Ads in an e-commerce system are commonly ranked by expected revenue, which can be estimated using a model trained on a pointwise loss function. It is often possible to obtain a better ranking using a model trained on a pairwise loss function, but the scores output by such a model are individually meaningless. To improve the ranking of ads, we built a reranker in the form of a lightweight logistic regression model, which uses as features the score from a pointwise model and the rank of the item as judged by a pairwise model. This approach improved live CTR and PTR without sacrificing ad revenue, while remaining compatible with strict production latency constraints.

Keywords

Sponsored search, ranking for sponsored products, re-ranking

1. Introduction

In large-scale ad marketplaces, a system that ranks items according to expected revenue typically requires well-calibrated estimates of click or sale probabilities; i.e.,

$$\text{expected revenue} = \Pr(\text{sale}) \times \text{revenue per sale}.$$

Done poorly, attempting to maximize revenue may undermine long-term user engagement by surfacing low-quality items with high bids. It is therefore critical for the model to be able to accurately discriminate between items.

Typically, pairwise (and listwise) rankers are better at ranking items than pointwise ones [1]. However, since pairwise loss functions such as the one used in RankNet [2] are invariant under monotonic transformations of the scores, the resulting scores are not directly interpretable as predicted sale probabilities; furthermore, the scores are liable to shift whenever the model is retrained [3]. Thus, to treat pairwise model scores as though they represented well-calibrated sale probabilities carries significant business risk. Still, the pairwise model captures item quality signals that generalize to ads ranking.

Ideally, we would like to attain the desirable properties of both models simultaneously: the well-calibrated scores of a pointwise model, and the discriminative ability of a pairwise model. Any solution must respect the fact that our setting is a large-scale e-commerce ad ranking system, where strict latency, stability, and governance constraints limit the class of deployable models.

One potential approach is to add a post-processing calibration step to the pairwise model, such as Platt scaling [4] or isotonic regression. However, the scores output by our pairwise model depend heavily on the query context: the same score in different contexts may correspond to very different sale probabilities. Any calibration layer would therefore need to account for the query context as well.

Another approach is to train a model on a multi-objective function (e.g., a linear combination of pointwise and pairwise loss functions). Although this appears to be a promising avenue for future exploration, we chose a simpler approach that would take advantage of our existing models.

We implemented a reranker[1] that rescores the top ~100 items (as scored by the current pointwise model) for a given query, using as features the outputs of the pointwise and pairwise models. Specifically,

2026 SIGIR Workshop on eCommerce, July 24, 2026, Melbourne, Australia

ehallman@ebay.com (E. Hallman)

0000-0001-7908-2296 (E. Hallman)

[1]Loosely, meaning a model that is run when the number of selected items is small enough to admit sequential computation instead of parallel, and whose scores for a given item may depend on the whole set of selected items.

the reranker uses a logistic regression model of the form

$$\text{logit}(p_{\text{final}}) = \alpha \cdot \text{logit}(p_{\text{init}}) + \beta_r,$$

where p_{init} is the score from the pointwise model, and r is the item's rank (according to the pairwise model) among the set of candidates for reranking. The reranker model was fit using the standard pointwise log-loss objective function.

The idea to combine scores from separate models was not new. Previously, we had been ranking items according to the fusion model

$$\text{ranking score} = (s_{\text{pointwise}})^{\alpha} \cdot (s_{\text{pairwise}})^{1-\alpha} \tag{1}$$

where α was a tuneable parameter. Although this model gave better rankings than the pointwise model alone, it had the two disadvantages mentioned earlier: the ranking score was not a calibrated estimate of either sale probability or eCPM, and it was sensitive to changes in the pairwise ranker. The reranker described in this work mitigated these disadvantages without sacrificing performance.

In offline analysis, the reranker model showed 0.3% higher ROC AUC than the pointwise model. It was later tested in combination with a separate update to the pointwise model. We found that when we replaced the old fusion model (1) with ranking purely by the updated pointwise model, it increased ad revenue at the expense of user metrics (CTR, PTR, mean reciprocal rank of first sale, etc.). By contrast, the combination of the updated pointwise model and the new reranker showed significant improvements in all of these metrics without sacrificing ad revenue. Improvements were concentrated in the highest-ranked items, which dominate auction outcomes.

1.1. Related work

The general idea of combining scores from multiple models, known as *data fusion* or *rank aggregation* [5], has been studied for decades; see [6] and the many references therein. Our approach has some relation to *ProbFuse* [7] in that it maps ranks to scores based on training data. One notable difference is that our model uses a hybrid of a pointwise estimate from one model and a ranking from another. Another difference is that for *ProbFuse* and its variants the scoring function is presented as a sum of probabilities over the models; our approach, by contrast, is approximately multiplicative:

$$p_{\text{final}} \approx p_{\text{init}}^{\alpha} \cdot \exp(\beta_r).$$

When $\alpha = 1$ the rank-based terms β_r can be interpreted as applying a log-likelihood correction to the pointwise predictions p_{init}. Lillis et al. [8] state that adding the probability scores is generally more useful for data fusion problems, but acknowledge that multiplying the probabilities would be an "obvious" choice. Although the ranking scores output by *ProbFuse* are not to be directly interpreted as probabilities, Anava et al. [6] have developed a probabilistic framework that could be used to explain it and other fusion methods.

Some works that study model calibration through a post-processing step include [9, 10], which use Platt scaling or a similar correction, [11], which introduces Beta calibration, and [12], which considers a few other methods including isotonic and polynomial regression.

In recent years, significant attention has been directed toward training a model on a multi-objective loss function. IntervalRank [13] includes information from an ordinal variable representing the item's relevance score, so that items with equal relevance judgments get similar scores and items with different judgments have well-separated scores. [14] examines ranking distillation, where the loss function is a combination of ranking loss and the deviation from the scores of a teacher model. [15] finds that training on a mixed-objective function can lower the pointwise loss on both training and validation sets, and suggests that this is because the incorporation of the pairwise loss prevents the gradients from vanishing on negative samples. One interesting subset of these works has explored *regression compatible* or *calibration compatible* loss functions [1, 16], where the individual losses in the multi-objective function can be minimized simultaneously. For more references, see the background given in [15].

More loosely related to this work is the field of order-restricted statistical inference [17]. If the true ordering of a set of target values $\{y_i\}_{i=1}^n$ is known, the general solution is to fit an isotonic model

$$\hat{\mu} = \underset{\mu_1 \geq \ldots \geq \mu_n}{\arg\min} \sum_{i=1}^{n} w_i \cdot \ell(y_i - \mu_i) \tag{2}$$

for some weights w_i and loss function $\ell(\cdot)$. For a least-squares loss function the problem can be solved using the pool adjacent violators algorithm [18]. In our case, the pairwise model might be considered as a noisy oracle that provides a superior but still-imperfect ranking compared to the pointwise model. A more appropriate model might then be *nearly-isotonic regression* [19], where the hard inequality constraints in (2) are replaced with a soft penalty function. One potential model might then be

$$\hat{\mu} = \underset{\mu_1,\ldots,\mu_n}{\arg\min} \frac{1}{2} \sum_{i=1}^{n} (s_i - \mu_i)^2 + \lambda \sum_{i=1}^{n-1} (\mu_i - \mu_{i+1})_+,$$

where the pointwise logit scores s_i are ordered according to the oracle, the first sum encourages alignment with the pointwise scores, and the second sum encourages alignment with the oracle's ordering. Further exploration of this approach is outside the scope of this work.

2. Methods

We faced two main design choices for this problem: what class of models to consider when training the reranker model, and what loss function to use.

2.1. Choice of model

We trained a logistic regression model of the form

$$\text{logit}(p_{\text{final}}) = \alpha \cdot \text{logit}(p_{\text{init}}) + \beta_r + \gamma_{\text{slot}}, \tag{3}$$

where p_{init} is the score from the pointwise model, and r is the item's rank among the recall set according to the pairwise model. The weight γ_{slot} accounts for the slot on the search results page in which the item appeared in training data, and is not used at inference time. Because the slot is strongly correlated with the pairwise rank r, we first estimated the slot biases γ_{slot}, then treated these weights as fixed when fitting α and β_r in (3). This model could be interpreted as a generalization of Platt scaling, permitting a different intercept for each value of r. In order to make the model more robust, the rank r was clamped to a maximum of 20 both during training and at inference time.

One notable property of this model is that it uses only the *rank* of each item from the pairwise model, and not the raw score. Since the rank can be computed only after aggregating the items for a given query, our model must be implemented as a reranker. We also expect this model to be robust to retraining the pairwise model, since the ranking determined by the pairwise model will be far more stable over time than the raw scores.

2.1.1. Calibration

We implemented a layer of calibration between the pointwise model and the reranker model: the inputs p_{init} in (3) are the post-calibration scores. We expect that maintaining this calibration layer will make the reranker model less sensitive to updates to the pointwise model.

No further layer of calibration was added after the reranker.

2.1.2. Other possible models

In practice, we observed that the weights β_r were mostly, but not entirely, monotonically decreasing in r. Since the weights should intuitively be monotonic, this suggests that the model may be overfitting the training data. We do not consider this to be cause for much concern, because the logistic regression model has very few parameters and because the deviation from monotonicity is slight. We could, however, bin the ranks r more coarsely, or simply require the weights β_r to be monotonic when training the model.

We considered using a few other rank-based inputs, such as the item's rank when ordered by price from cheapest to most expensive. In offline simulations we found that adding more features led to a minor increase in ROC AUC, but in order to simplify the implementation we used only the rank from the pairwise model in online tests.

We considered but have not yet tested enforcing $\alpha = 1$, which could lead to better-calibrated predictions for extreme values of p_{init}. In the opposite direction, we could also make the model more flexible, though perhaps less robust, by allowing α to vary with r.

A team at Google [9] found that compared to a variant of Platt scaling, piecewise linear isotonic regression "significantly reduced bias" at the extreme ends of the distribution. We could therefore also consider fitting an isotonic regression model for each value of r, or fitting a bivariate isotonic regression model [20].

2.2. Choice of loss function

We trained the reranker model with the standard pointwise cross-entropy loss function. The training data consisted of ad items appearing on the first page of search results. The target label was whether an item's sale was credited to that search, according to the company's attribution logic. We used negative downsampling (taking 10% of negative samples) to increase the proportion of positive labels, and weighted the negative samples accordingly for the training loss.

We considered one other loss function: the same pointwise cross-entropy loss, but weighted by each item's potential ad revenue. We trained the model using weights $w_i = 1 + \log(1 + \text{revenue}_i)$, but found no notable difference in offline analysis from the unweighted version and did not test this weighted version on live traffic.

In the future we may test more variants that weight the loss by ad revenue. We may also consider mixed objective functions such as the ones discussed in [1] or [16].

3. Experimental Results

For offline analysis, we trained the reranker model on items appearing on the first page of search results, and tested it on items appearing in the first ad slot. Compared to the pointwise model alone, the re-ranked scores showed a +0.27% increase in ROC AUC.

A simpler approach, such as using isotonic regression on the output of the pairwise model, was not seriously considered because the pairwise model had significantly worse AUC (more than 10% lower) over the test dataset. This finding does not contradict the superior ranking ability of the pairwise model because its scores can depend on the query context.

We assessed the calibration error of offline results by sorting items by the model score and splitting into 10 bins. The 2-norm of the calibration error across the 10 bins was comparable between the pointwise and reranker models, but the mean absolute percentage error (MAPE) was about 35% higher for the reranker model. Although the increase in MAPE merits further study, we emphasize that using the pointwise model alone was not considered an acceptable option due to the resulting degradation to user metrics (see Table 1 for live traffic results). The relevant point of comparison is to the former fusion model (1), which did not rank items according to calibrated scores at all.

Table 1
Effect of adding the reranker as compared to scoring by the pointwise model alone. Here, R refers to revenue-related metrics, and U refers to user experience metrics. When updating the pointwise model alone, all changes to user metrics were directionally bad; including the reranker mitigated or fully reversed these changes.

Metric	Description	Pointwise model	With reranker	Change
R1	Ad revenue	**+1.35%** ±0.43%	**+1.52%** ±0.37%	+0.17%
R2	Ad revenue	**+0.43%** ±0.65%	**+1.63%** ±0.54%	+1.20%
R3	Ad revenue	+0.94% ±1.04%	**+4.06%** ±0.85%	+3.12%
R4	Site revenue	+0.25% ±0.39%	+0.33% ±0.34%	+0.08%
R5	Site revenue	+0.09% ±0.46%	+0.07% ±0.40%	-0.02%
U1	Engagement	**+6.78%** ±0.92%	**+2.92%** ±0.71%	-3.86%
U2	Engagement	**-0.77%** ±0.17%	**+2.76%** ±0.15%	+3.53%
U3	Engagement	**-5.45%** ±0.58%	**+1.48%** ±0.50%	+6.93%
U4	Ranking	**-1.10%** ±0.28%	**+0.47%** ±0.24%	+1.57%

3.1. Test on live traffic

We tested the reranker in combination with an update to the pointwise model in a 3-week A/B test. We therefore do not have results that show the effects of the reranker in isolation — the closest data we have for comparison is a 2-week experiment that replaced the old fusion model (1) with the pointwise model, without the reranker. Still, the results are striking enough that we consider them to be informative.

The differences in the two test results are shown in Table 1. For both tests, the control group used the old fusion model, the ranking scores of which were not calibrated sale probabilities. On its own, the updated pointwise model put much more expensive items at the top of the search results, attaining an overall increase in ad revenue at the expense of both clicks and sales. With the reranker added, the items being shown were still more expensive compared to the control version, but user metrics including clicks, sales, and the mean reciprocal rank of the first sale all improved.

Interestingly, neither treatment had a statistically significant effect on site-wide revenue or GMB[2]. This suggests that one of the main contributions of the reranker may be to more highly value high-quality items in organic slots, and to put them in ad slots instead.

4. Conclusions and Future Work

By implementing a reranker that uses as inputs the results of two models, one trained on a pointwise loss function and one on a pairwise loss function, we were able to significantly improve user metrics (including CTR and PTR) on live traffic for a large e-commerce system. We consider it interesting that significant improvements were possible using only a simple logistic regression model for the reranker, and by considering only the local rank assigned by the pairwise model rather than its distribution of scores.

In future work we plan to test a calibration layer after the reranker, and to experiment with other choices for the loss function. The latter effort could include weighting the training samples by revenue, or could involve a more significant change such as adopting a multi-objective function. We will also consider changes to the model itself, such as requiring the weights to be monotonic with respect to the pairwise rank, or fitting a separate isotonic regression model for each such rank.

Acknowledgments

The author thanks Sanjana Arun, Greg Kocher, and Yaojia Zhu for their helpful comments, and Angel Rodriguez for his work in implementing the reranker.

[2]Gross Merchandise Bought; i.e, the total value of all purchased items.

Declaration on Generative AI

The author did not employ any Generative AI tools in writing the first draft of this article. During the revision process, an LLM was consulted in locating background references on data fusion. The choice of references, and the summaries of these works, are the author's alone.

References

[1] Aijun Bai, Rolf Jagerman, Zhen Qin, Le Yan, Pratyush Kar, Bing-Rong Lin, Xuanhui Wang, Michael Bendersky, Marc Najork, Regression Compatible Listwise Objectives for Calibrated Ranking with Binary Relevance, in: Proceedings of the 32nd ACM Conference on Information and Knowledge Management, CIKM '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 4502–4508. doi:10.1145/3583780.3614712.

[2] Chris J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, Greg Hullender, Learning to Rank using Gradient Descent, Technical Report MSR-TR-2005-06, Microsoft Research, 2005. URL: https://www.microsoft.com/en-us/research/publication/learning-to-rank-using-gradient-descent/.

[3] Le Yan, Zhen Qin, Xuanhui Wang, Michael Bendersky, Marc Najork, Scale Calibration of Deep Ranking Models, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 4300–4309. doi:10.1145/3534678.3539072.

[4] John Platt, Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods, Advances in Large Margin Classifiers 10 (1999) 61–74.

[5] D. Lillis, On the evaluation of data fusion for information retrieval, in: Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '20, Association for Computing Machinery, New York, NY, USA, 2021, p. 54–57. URL: https://doi.org/10.1145/3441501.3441506. doi:10.1145/3441501.3441506.

[6] Y. Anava, A. Shtok, O. Kurland, E. Rabinovich, A probabilistic fusion framework, in: S. Mukhopadhyay, C. Zhai, E. Bertino, F. Crestani, J. Mostafa, J. Tang, L. Si, X. Zhou, Y. Chang, Y. Li, P. Sondhi (Eds.), Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016, ACM, 2016, pp. 1463–1472. URL: https://doi.org/10.1145/2983323.2983739. doi:10.1145/2983323.2983739.

[7] D. Lillis, F. Toolan, R. Collier, J. Dunnion, ProbFuse: A Probabilistic Approach to Data Fusion, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06), ACM, Seattle, WA, USA, 2006, pp. 139–146. doi:10.1145/1148170.1148197.

[8] D. Lillis, L. Zhang, F. Toolan, R. W. Collier, D. Leonard, J. Dunnion, Estimating Probabilities for Effective Data Fusion, in: Proceedings of the 33rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Geneva, Switzerland, 2010, pp. 347–354. doi:10.1145/1835449.1835508.

[9] H. Brendan McMahan, Gary Holt, D. Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos, Jeremy Kubica, Ad Click Prediction: a View from the Trenches, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 1222–1230. doi:10.1145/2487575.2488200.

[10] Yukihiro Tagami, Shingo Ono, Koji Yamamoto, Koji Tsukamoto, Akira Tajima, CTR Prediction for Contextual Advertising: Learning-to-Rank Approach, in: Proceedings of the Seventh International Workshop on Data Mining for Online Advertising, ADKDD '13, Association for Computing Machinery, New York, NY, USA, 2013. doi:10.1145/2501040.2501978.

[11] Meelis Kull, Telmo Silva Filho, Peter Flach, Beta Calibration: a Well-Founded and Easily Imple-

mented Improvement on Logistic Calibration for Binary Classifiers, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, volume 54 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 623–631. URL: https://proceedings.mlr.press/v54/kull17a.html.

[12] Sougata Chaudhuri, Abraham Bagherjeiran, James Liu, Ranking and Calibrating Click-Attributed Purchases in Performance Display Advertising, in: Proceedings of the ADKDD'17, ADKDD'17, Association for Computing Machinery, New York, NY, USA, 2017. doi:10.1145/3124749.3124755.

[13] Taesup Moon, Alex Smola, Yi Chang, Zhaohui Zheng, IntervalRank: Isotonic Regression with Listwise and Pairwise Constraints, in: Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10, Association for Computing Machinery, New York, NY, USA, 2010, p. 151–160. doi:10.1145/1718487.1718507.

[14] Jiaxi Tang, Ke Wang, Ranking Distillation: Learning Compact Ranking Models With High Performance for Recommender System, in: Proceedings of the 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 2289–2298. doi:10.1145/3219819.3220021.

[15] Zhutian Lin, Junwei Pan, Shangyu Zhang, Ximei Wang, Xi Xiao, Shudong Huang, Lei Xiao, Jie Jiang, Understanding the Ranking Loss for Recommendation with Sparse User Feedback, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 5409–5418. doi:10.1145/3637528.3671565.

[16] Xiaoqiang Gui, Yueyao Cheng, Xiang-Rong Sheng, Yunfeng Zhao, Guo-Xian Yu, Shuguang Han, Yuning Jiang, May Xu, Bo Zheng, Calibration-Compatible Listwise Distillation of Privileged Features for CTR Prediction, in: WSDM '24: The 17th ACM International Conference on Web Search and Data Mining, 2024, pp. 247–256. doi:10.1145/3616855.3635810.

[17] Tim Robertson, Farroll T. Wright, Richard Dykstra, Order Restricted Statistical Inference, Wiley Series in Probability and Mathematical Statistics, Wiley, Chichester, 1988. Bibliography: p. [459]-508.

[18] Richard E. Barlow, Statistical Inference Under Order Restrictions: the Theory and Application of Isotonic Regression, Wiley Series in Probability and Mathematical Statistics, no. 8, J. Wiley, London, 1972. Bibliography: p. 365–377.

[19] Ryan Tibshirani, Holger Höfling, Robert Tibshirani, Nearly-Isotonic Regression, Technometrics 53 (2011). doi:10.2307/40997292.

[20] S. Sasabuchi, M. Inutsuka, D. D. Sarath Kulatunga, A Multivariate Version of Isotonic Regression, Biometrika 70 (1983) 465–472. doi:10.1093/biomet/70.2.465.